**K-MEANS CLUSTERING**

Let $x_1, \ldots, x_n \in \mathbb{R}^d$.

Recall that the goal of clustering is to assign the data to disjoint subsets called ________ so that points in the same cluster are more similar than points in different clusters.

Therefore, at the heart of every clustering algorithm is a notion of ________.

Often it is more convenient to work with a ________.
Dissimilarity

A dissimilarity matrix is an $n \times n$ matrix

$$D = \left[ d_{ij} \right]_{i,j=1}^n$$

which has the following properties

- $d_{ii} = 0$
- $d_{ij} = d_{ji}$
- $d_{ij} \geq 0$

Conceptually, if $x_i$ is more similar to $x_j$ than to $x_k$, then

A dissimilarity matrix need not satisfy the triangle inequality:
Examples

- Euclidean distance
- Squared Euclidean distance
- kNN-based distance
A cluster map is a function

\[ C: \{1, 2, \ldots, n\} \rightarrow \{1, 2, \ldots, K\} \]

that partitions the data into \( K \) clusters.

In K-means clustering, we

- assume \( K \) is known (more on this later)
- adopt the squared Euclidean distance as a dissimilarity

\[ d_{ij} = \]

- seek to minimize the \[ W(c) = \]

where

\[ n_k = \]
The $K$-means criterion is an optimization problem. The number of possible cluster maps $C$ is

$$\frac{1}{K!} \sum_{k=1}^{K} (-1)^{K-k} \binom{K}{k} k^n$$  (Jain & Dubes, 1988)

$$\left\{ \begin{array}{ll}
= 34,105 & \text{if } n = 10, \ K = 4 \\
\approx 10^{10} & \text{if } n = 19, \ K = 4
\end{array} \right.$$

There is no known efficient search strategy for this space. Therefore we resort to an iterative, suboptimal algorithm.
Exercise 1  Show that

\[ W(C) = \sum_{k=1}^{K} \sum_{i: C(i) = k} \| x_i - \overline{x}_k \|^2 \]

where

\[ \overline{x}_k := \frac{1}{n_k} \sum_{i: C(i) = k} x_i \]
Solution

\[
W(c) = \frac{1}{2} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i: C(i) = k} \sum_{j: C(j) = k} \left\| x_i - \bar{x}_k - (x_j - \bar{x}_k) \right\|^2
\]

\[
\left< x_i - \bar{x}_k - (x_j - \bar{x}_k), x_i - \bar{x}_k - (x_j - \bar{x}_k) \right>
\]

\[
= \left\| x_i - \bar{x}_k \right\|^2 - 2 (x_i - \bar{x}_k)^T (x_j - \bar{x}_k) + \left\| x_j - \bar{x}_k \right\|^2
\]

\[
= \frac{1}{2} \sum_{k=1}^{K} \frac{1}{n_k} \left[ \sum_{i: C(i) = k} \sum_{j: C(j) = k} \left\| x_i - \bar{x}_k \right\|^2
\]

\[
- 2 \sum_{i: C(i) = k} \sum_{j: C(j) = k} (x_i - \bar{x}_k)^T (x_j - \bar{x}_k)
\]

\[
+ \sum_{i: C(i) = k} \sum_{j: C(j) = k} \left\| x_j - \bar{x}_k \right\|^2 \right]
\]

\[
= \frac{1}{2} \sum_{k=1}^{K} \frac{1}{n_k} \left[ n_k \sum_{i: C(i) = k} \left\| x_i - \bar{x}_k \right\|^2
\]

\[
+ n_k \sum_{j: C(j) = k} \left\| x_j - \bar{x}_k \right\|^2 \right]
\]

\[
= \sum_{k=1}^{K} \sum_{i: C(i) = k} \left\| x_i - \bar{x}_k \right\|^2
\]
Figure 2-1-1 (A): LSMDE for $N(0,1)$ with positive weights ($n=100$). The solid line is LSMDE and the dotted line is $N(0,1)$.
1) \( m_k^* = \)

2) \( C^*(i) = \)

K-means Clustering Algorithm

Initialize \( \bar{x}_k, \ k=1,...,K \)

Repeat

• \( C(i) = \)

• \( \bar{x}_k = \)

Until clusters don’t change

Remarks

• The algorithm is typically initialized by setting each \( \bar{x}_k \) to be a random data point

• Since the algorithm often finds a local min, several random initializations are recommended.
Cluster Geometry

Clusters are "nearest neighbors"

regions or _____ cells defined with respect to the cluster means.

Therefore the cluster boundaries are

\[ K = 3 \]

\[ k \text{-means will fail if clusters are} \]
Model selection

How to choose $K$?

If $W_k(C^*)$ is the within-cluster scatter based on $K$ clusters, we have a plot like this:

![Plot](image)

If the "right" number of clusters is $K^*$, we expect:

1. For $K < K^*$, $W_k(C^*) - W_{k-1}(C^*)$ will be large.
2. For $K > K^*$, $W_k(C^*) - W_{k-1}(C^*)$ will be small.

This suggests choosing $K$ near the "knee" of the curve.
A. clusters, similarity, dissimilarity

B. \( d_{ij} < d_{ik} \)
\[ d_{ij} + d_{jk} \neq d_{ik} \]

C. \( \| \mathbf{x} - \mathbf{y} \| = \left( \sum_{j=1}^{d} (x(j) - y(j))^2 \right)^{\frac{1}{2}} \)
\( \| \mathbf{x} - \mathbf{y} \|^2 \)
- length of shortest path on \( k \)-nearest neighbor graph, for some \( k \)

D. \( d_{ij} = \| x_i - x_j \|^2 \), within cluster scatter
\[ W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{i:C(i)=k} \left[ \frac{1}{n_k} \sum_{j:C(j)=k} \| x_i - x_j \|^2 \right] \]
- avg. dissim. to points in same cluster

\[ n_k = \frac{\sum_{i=1}^{n} 1 \{ C(i) = k \}}{n} \]

E. combinatorial
F. \( \bar{x}_k \)
\[ n_k^* = \frac{1}{n_k} \sum_{i : C(i) = k} x_i \]

\[ C^*(i) = \arg \min_k ||x_i - \bar{x}_k|| \]

\[ C(i) = \arg \min_k ||x_i - \bar{x}_k|| \]

\[ \bar{x}_k = \frac{1}{n_k} \sum_{i : C(i) = k} x_i \]

I. Voronoi, hyperplanes, nonconvex